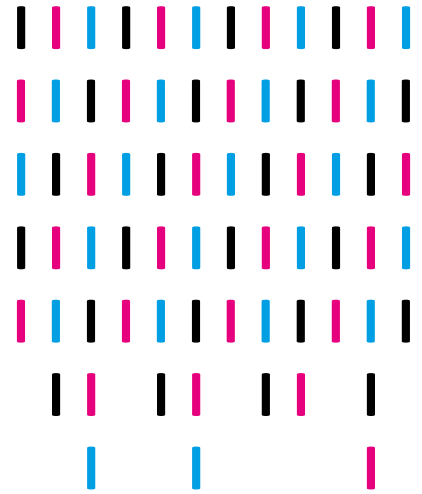




NEPS *SURVEY PAPERS*

Timo Gnambs

NEPS TECHNICAL REPORT FOR VERBAL AND NONVERBAL REASONING: SCALING RESULTS OF STARTING COHORT 8 FOR GRADE 5 IN SPECIAL SCHOOLS

NEPS Survey Paper No. 119
Bamberg, April 2025

Survey Papers of the German National Educational Panel Study (NEPS)

at the Leibniz Institute for Educational Trajectories (LIfBi) at the University of Bamberg

The NEPS *Survey Paper* series provides articles with a focus on methodological aspects and data handling issues related to the German National Educational Panel Study (NEPS).

They are of particular relevance for the analysis of NEPS data as they describe data editing and data collection procedures as well as instruments or tests used in the NEPS survey. Papers that appear in this series fall into the category of 'grey literature' and may also appear elsewhere.

The NEPS *Survey Papers* are edited by a review board consisting of the scientific management of LIfBi and NEPS.

The NEPS *Survey Papers* are available at www.neps-data.de (see section "Publications") and at www.lifbi.de/publications.

Editor-in-Chief: Thomas Bäumer, LIfBi

Review Board: Board of Directors, Heads of LIfBi Departments, and Scientific Management of NEPS Working Units

Contact: German National Educational Panel Study (NEPS) – Leibniz Institute for Educational Trajectories – Wilhelmsplatz 3 – 96047 Bamberg – Germany – contact@lifbi.de



This work is licensed under CC BY 4.0. To view a copy of this license, visit <https://creativecommons.org/licenses/by/4.0/>

NEPS Technical Report for Verbal and Nonverbal Reasoning: Scaling Results of Starting Cohort 8 for Grade 5 in Special Schools

Timo Gnambs 

Leibniz Institute for Educational Trajectories

E-mail address of the lead author:

timo.gnambs@lifbi.de

Bibliographic data:

Gnambs, T. (2025). *NEPS technical report for verbal and nonverbal reasoning: Scaling results of Starting Cohort 8 for grade 5 in special schools* (NEPS Survey Paper No. 119). Leibniz Institute for Educational Trajectories, National Educational Panel Study. <https://doi.org/10.5157/NEPS:SP119:1.0>

Acknowledgements. This report is an extension to *NEPS Survey Paper 118* (Gnambs, 2025) that presents the scaling results for reading in Grade 5 of Starting Cohort 8. Therefore, various parts of this report (e.g., regarding the introduction and the analytic strategy) are reproduced verbatim to facilitate the understanding of the presented results.

NEPS Technical Report for Verbal and Nonverbal Reasoning: Scaling Results of Starting Cohort 8 for Grade 5 in Special Schools

Abstract

The National Educational Panel Study (NEPS) examines the development of competencies across the lifespan. To evaluate the quality of the tests administered in the NEPS, several analyses based on item response theory are conducted. This paper describes the data and scaling procedures for tests of verbal and nonverbal reasoning abilities that was administered in Grade 5 of Starting Cohort 8 in special schools. Each subscale consisted of 25 items that were administered to 212 students (37% girls) from special schools. A Rasch model was used to scale the data. Item fit statistics, Rasch-homogeneity, and local item independence were evaluated to ensure the quality of the test. The analyses showed that the items had satisfactory item fit and good reliability that allowed the estimation of reliable person scores. Furthermore, analyses of differential item functioning showed comparable measurement models for the test administered in special schools of Starting Cohort 3 and Starting Cohort 8. In addition to the scaling results, this report also describes the data available in the scientific use file.

Keywords

item response theory, scaling, verbal reasoning, nonverbal reasoning, scientific use file

Contents

1	Introduction	4
2	Testing Verbal and Nonverbal Reasoning	4
3	Sample	5
4	Psychometric Analyses	5
4.1	Missing Responses	5
4.2	Scaling Model	5
4.3	Software	6
5	Results	6
5.1	Verbal Reasoning	6
5.1.1	Missing Responses	6
5.1.2	Parameter Estimates	10
5.1.3	Quality of the test	14
5.2	Nonverbal Reasoning	16
5.2.1	Missing Responses	16
5.2.2	Parameter Estimates	19
5.2.3	Quality of the test	23
6	Discussion	25
7	Data in the Scientific Use Files	25
7.1	Naming Conventions	25
7.2	Cognitive Ability Scores	25
	References	27

1. Introduction

The National Educational Panel Study (NEPS) measures different competencies coherently across the lifespan. These include, among others, reading competence, mathematical competence, and domain-general cognitive functioning. An overview of the competencies measured in the NEPS is given by Weinert et al. (2011) and Fuß et al. (2025). Most of the administered tests are routinely evaluated using psychometric models based on item response theory (IRT, see Pohl & Carstensen, 2012).

This report presents the results of these analyses for tests of verbal and nonverbal reasoning abilities that were administered in fifth grades of special schools in Starting Cohort 8 (fifth grade). The following sections first introduce the main concepts of the administered tests and the test design. Then, the sample is described and the results of various psychometric analyses are reported that were conducted to check the quality of the test. Finally, an overview of the data that are available for public use in the scientific use file (SUF) is given.

The analyses summarized in this report are based on the data available at some point prior to the public data release. Due to ongoing data protection and data cleansing issues, the data in the SUF may differ slightly from the data used for the analyses in this report. However, pronounced differences in the presented results are not expected for the final data available in the SUF.

2. Testing Verbal and Nonverbal Reasoning

Verbal and nonverbal reasoning were assessed with two subscales of the *Cognitive Abilities Test* (Heller & Perleth, 2000), a German version of Thorndike's cognitive abilities test (Thorndike & Hagen, 1971), that measures acquired reasoning and problem-solving abilities relevant for learning at school. The subscale *vocabulary* measuring verbal reasoning included 25 items, each requiring to identify one out of five words corresponding to a written target word (e.g., synonyms). However, Item 20 was not considered in the present analyses because it yielded a negative discrimination in preliminary analyses (see Gnambs & Nusser, 2024, for similar findings in Starting Cohort 3). The subscale *figure classifications* measuring nonverbal reasoning included 25 items, each presenting three or four figures that could be classified according to a distinct characteristic (e.g., form, shading, position). The students had to identify one out of five figures that matched the classification of the target stimuli. Because these tests were developed for general student populations, the study administered test versions that were designed for a younger age group from fourth grade (see Gnambs & Nusser, 2024).

The study was conducted in small groups at the respective schools of the students. The tests were presented on paper by trained test administrators. The maximum testing time was 7 and 9 minutes for the verbal and figural reasoning subtests, respectively. The study assessed different cognitive domains including reading competence (Gnambs, 2025) and reasoning abilities. The cognitive abilities test was always presented second after the reading test. There was no multi-matrix design with respect to the order of the items within the test. All respondents received the test items in the same order. A comprehensive description of the study design is available on the NEPS website.¹

¹<https://www.neps-data.de>

3. Sample

A total of 212 participants (37% girls) were administered the reasoning test. However, 3 participants had less than three valid responses to the items of the verbal reasoning subtest. Because no reliable competence scores can be estimated from such a small number of responses, these cases were excluded from the psychometric analyses (see Pohl & Carstensen, 2012). Therefore, the analyses for the verbal reasoning subtest were based on 209 respondents, while the analyses for the nonverbal reasoning subtest used 212 respondents.

4. Psychometric Analyses

4.1 Missing Responses

Competence data contain different types of missing responses. These are missing responses due to a) invalid responses, b) omitted items, and c) items that test takers did not reach. Invalid responses occurred, for example, when two response options were selected although only one was required. Omitted items occurred when respondents skipped some items. Due to time constraints or lack of motivation, not all participants completed the test within the allotted time. All missing responses after the last valid response were coded as not reached.

Missing responses provide information about how well the test worked (e.g., time limits, understanding of instructions, dealing with different response formats). Therefore, the occurrence of missing responses in the test was evaluated to get an idea of how well respondents coped with the test. Missing responses per item were examined to evaluate how well each of the items functioned.

4.2 Scaling Model

Item and person parameters were estimated using a Rasch (1960) model with Gauss-Hermite quadrature (21 nodes). In line with the scoring approach recommended by the test authors (Heller & Perleth, 2000), these analyses scored missing responses as incorrect. A detailed description of the scaling model can be found in Pohl and Carstensen (2012), while the data available in the SUF are described in Section 7.

The items of the two subtests consisted of one correct response option and several distractors (i.e., incorrect response options). The quality of the distractors within the multiple-choice items was examined to evaluate whether the distractors were predominantly chosen by respondents with a lower ability rather than by those who gave a correct response. We therefore calculated the point-biserial correlation between selecting an incorrect response option for a given item and the total correct score for the remaining items. Negative correlations indicate good distractors, while correlations between .00 and .05 were considered acceptable and correlations above .05 were considered problematic distractors (Pohl & Carstensen, 2012).

The fit of the dichotomous items to the Rasch model was evaluated using the WMNSQ statistic, the respective *t*-value, and the item characteristic curves (see Pohl & Carstensen, 2012). Items with a WMNSQ > 1.15 (*t*-value > |6|) were considered to have a noticeable item misfit, and items with a WMNSQ > 1.20 (*t*-value > |8|) were considered to have a considerable item misfit and their performance was investigated further. Correlations of the item score with the

corrected total score greater than .30 were considered as good, those greater than .20 as acceptable, and those less than .20 as problematic. The overall judgment of item fit was based on all fit indicators.

The subtests of the cognitive abilities test were also administered in special schools of Starting Cohort 3 (see Gnambs & Nusser, 2024). To examine whether the item properties were comparable between cohorts and, thus cognitive scores can be compared between the two cohorts, differential item functioning (DIF) was investigated. DIF was examined using a multigroup item response model, in which the main effects of the cohorts as well as the differential effects of the cohorts on item difficulty were modeled. Based on experience with preliminary data, we considered absolute differences in estimated difficulties between the cohorts that were greater than 1 logit as very strong DIF, absolute differences between 0.6 and 1 as considerable and worthy of further investigation, differences between 0.4 and 0.6 as small but not severe, and differences less than 0.4 as negligible DIF (Pohl & Carstensen, 2012). In addition, measurement invariance was examined by comparing the fit of a model that acknowledged DIF to a model that included only main effects and no DIF using Akaike's (1974) information criterion (AIC) and Bayesian information criterion (BIC, Schwarz, 1978).

The subtests of the cognitive abilities tests were scaled using the Rasch model because it preserves the weighting of the different aspects of the framework as intended by the test developers (Pohl & Carstensen, 2012). Nevertheless, Rasch-homogeneity is an assumption that may not hold for empirical data. To test the assumption of equal item discrimination parameters, a two-parametric item response model (Birnbaum, 1968) was also fitted to the data and compared to the Rasch model.

The dimensionality of the test was evaluated by examining the residuals of the Rasch model. Approximately zero-order correlations, as indicated by Yen's (1984) Q_3 , suggest unidimensionality. Because the Q_3 tends to be slightly negative in the case of locally independent items, we report the corrected Q_3 , which has an expected value of 0. According to Yen (1993), values of Q_3 falling below .20 indicate essential unidimensionality.

4.3 Software

The item response models were estimated with the *TAM* package version 4.2-21 (Robitzsch et al., 2024) in *R* version 4.4.1 (R Core Team, 2024).

5. Results

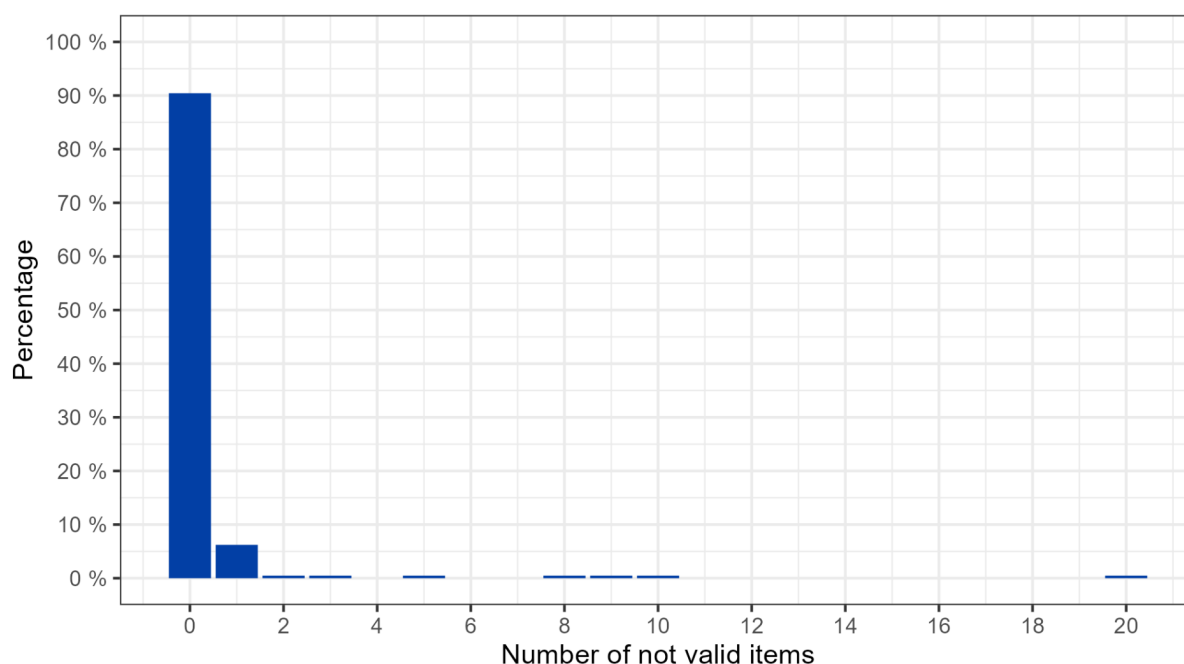
5.1 Verbal Reasoning

5.1.1 Missing Responses

5.1.1.1 Missing Responses per Person

The number of invalid responses per person is shown in Figure 1. The number of invalid responses was rather low with 90% of respondents having no invalid responses at all. Only about 3% of respondents had more than one invalid response.

Figure 2 illustrates the number of missing responses when respondents omitted items. The figure shows that a non-negligible number of items were omitted. Approximately 64% of re-

Figure 1*Number of Invalid Responses for Verbal Reasoning*

spondents had no omitted items, while 17% exhibited a single omitted item. However, less than 7% of respondents had five or more omitted items.

Another source of missing responses is items that respondents did not reach because they aborted the test, for example, due to time constraints or lack of motivation. These missing values refer to items after the last valid response. As shown in Figure 3, only about 54% of respondents did not abort the test and were administered all items. However, more than 93% of the sample received at least 8 items and thus about a third of the total test.

The total number of missing responses, aggregated across invalid, omitted, and not reached missing responses for each respondent, is shown in Figure 4. Because many respondents did not reach the end of the test or omitted some items, the total number of missing values was substantial. The median number of missing responses was 3; only about 35% of the respondents had no missing response at all. Almost 39% of respondents had more than five missing responses, while 21% had missing responses for 12 or more items (i.e., almost half of the administered items).

Overall, there was a small amount of invalid missing responses, while the percentage of omitted items was reasonable. However, the total number of missing responses was rather large because many respondents did not reach the end of the test.

5.1.1.2 Missing Responses per Item

Table 1 provides information on the occurrence of different types of missing responses per item. Overall, the number of respondents that omitted an item was acceptable. The percentage of omitted responses per item varied between 0% and 10% ($Mdn = 5\%$). In addition, most items

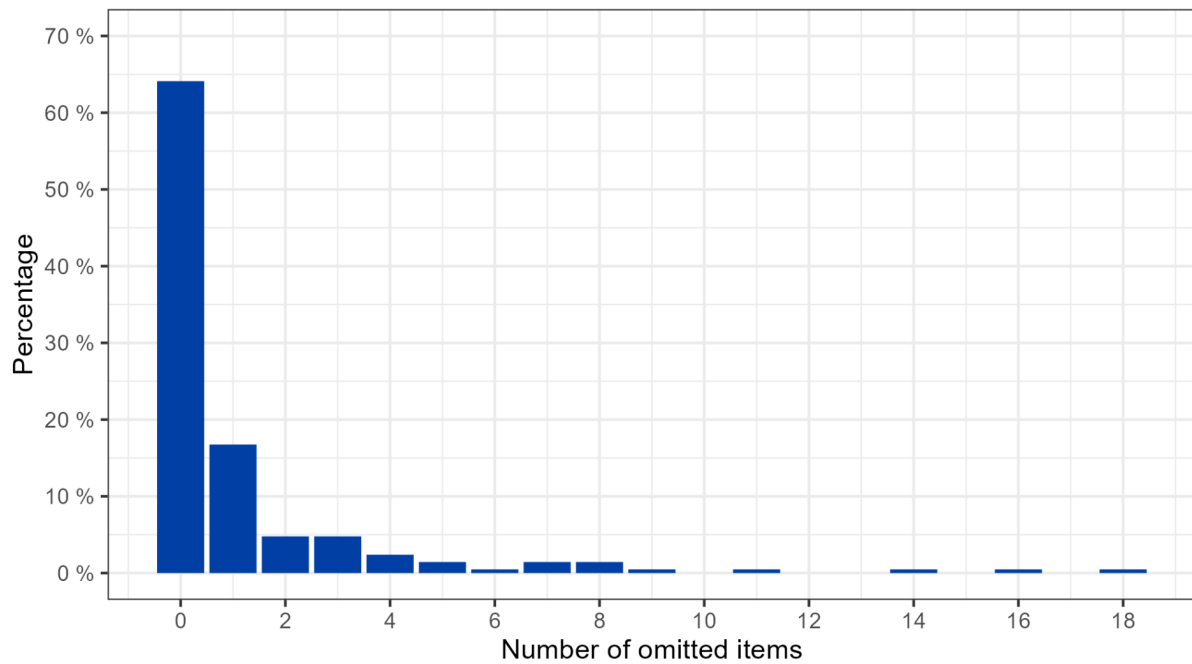
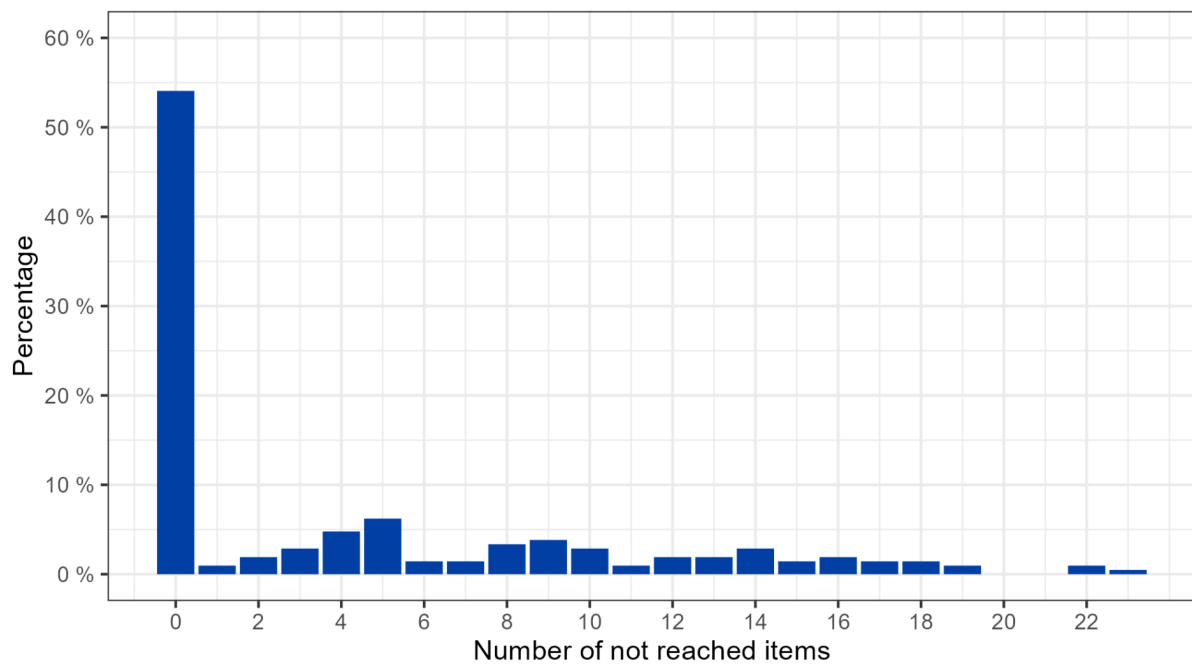
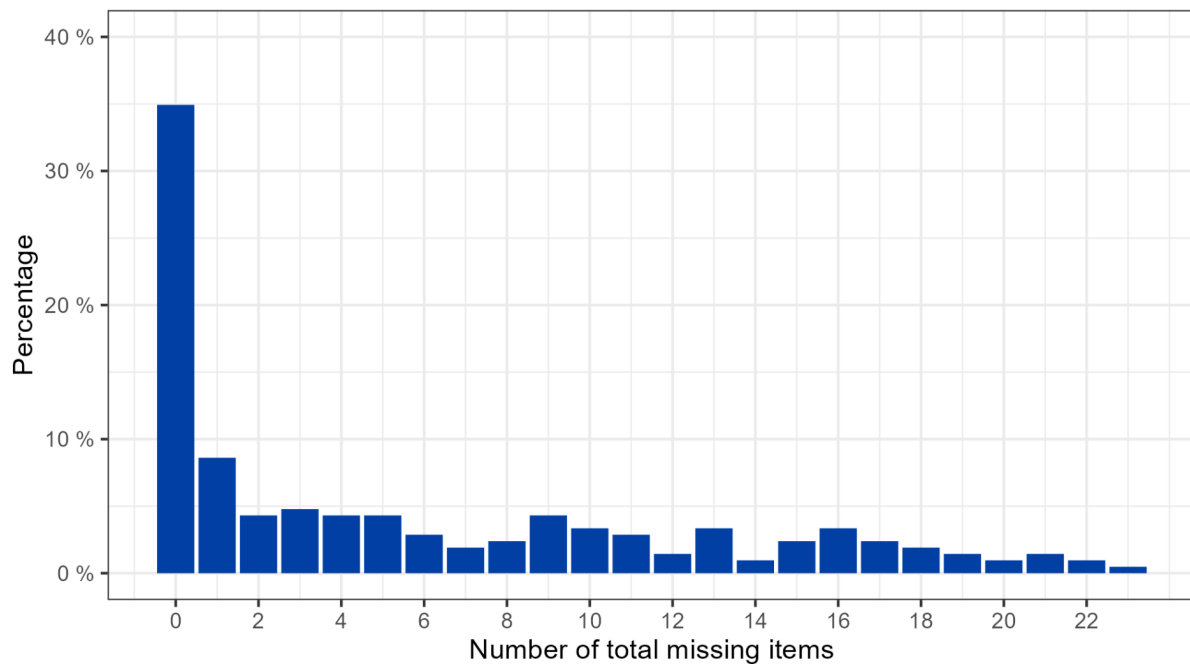
Figure 2*Number of Omitted Items for Verbal Reasoning***Figure 3***Number of Not-Reached Items for Verbal Reasoning*

Figure 4*Total Number of Missing Responses for Verbal Reasoning*

had relatively few invalid responses. The percentage of invalid responses varied across items between 0% and 3% ($Mdn = 1\%$).

Table 1*Percentage of Missing Responses by Item for Verbal Reasoning*

Nr.	Item	Pos.	N	OM	NR	NV
1	vig50101_sc8g5_c	1	199	2.39	0.00	2.39
2	vig50102_sc8g5_c	2	196	5.74	0.48	0.00
3	vig50103_sc8g5_c	3	184	8.61	1.44	1.91
4	vig50104_sc8g5_c	4	193	4.78	1.44	1.44
5	vig50105_sc8g5_c	5	191	5.26	1.44	1.91
6	vig50106_sc8g5_c	6	188	5.74	2.39	1.91
7	vig50107_sc8g5_c	7	187	3.83	3.83	2.87
8	vig50108_sc8g5_c	8	184	5.26	5.26	1.44
9	vig50109_sc8g5_c	9	174	8.13	7.18	1.44
10	vig50110_sc8g5_c	10	175	6.70	8.61	0.96
11	vig50111_sc8g5_c	11	176	2.87	11.48	1.44
12	vig50112_sc8g5_c	12	171	3.83	13.40	0.96

Nr.	Item	Pos.	N	OM	NR	NV
13	vig50113_sc8g5_c	13	162	5.26	15.31	1.91
14	vig50114_sc8g5_c	14	155	7.18	16.27	2.39
15	vig50115_sc8g5_c	15	146	10.05	19.14	0.96
16	vig50116_sc8g5_c	16	150	4.31	22.97	0.96
17	vig50117_sc8g5_c	17	142	4.78	26.32	0.96
18	vig50118_sc8g5_c	18	135	5.74	27.75	1.91
19	vig50119_sc8g5_c	19	139	3.35	29.19	0.96
20	vig50121_sc8g5_c	21	127	2.39	35.41	1.44
21	vig50122_sc8g5_c	22	115	3.35	40.19	1.44
22	vig50123_sc8g5_c	23	113	1.44	43.06	1.44
23	vig50124_sc8g5_c	24	107	3.35	44.98	0.48
24	vig50125_sc8g5_c	25	113	0.00	45.93	0.00

Note. Pos. = Item position within the test. N = Number of valid responses. NR = Percentage of respondents that did not reach an item. OM = Percentage of omitted responses. NV = Percentage of invalid responses.

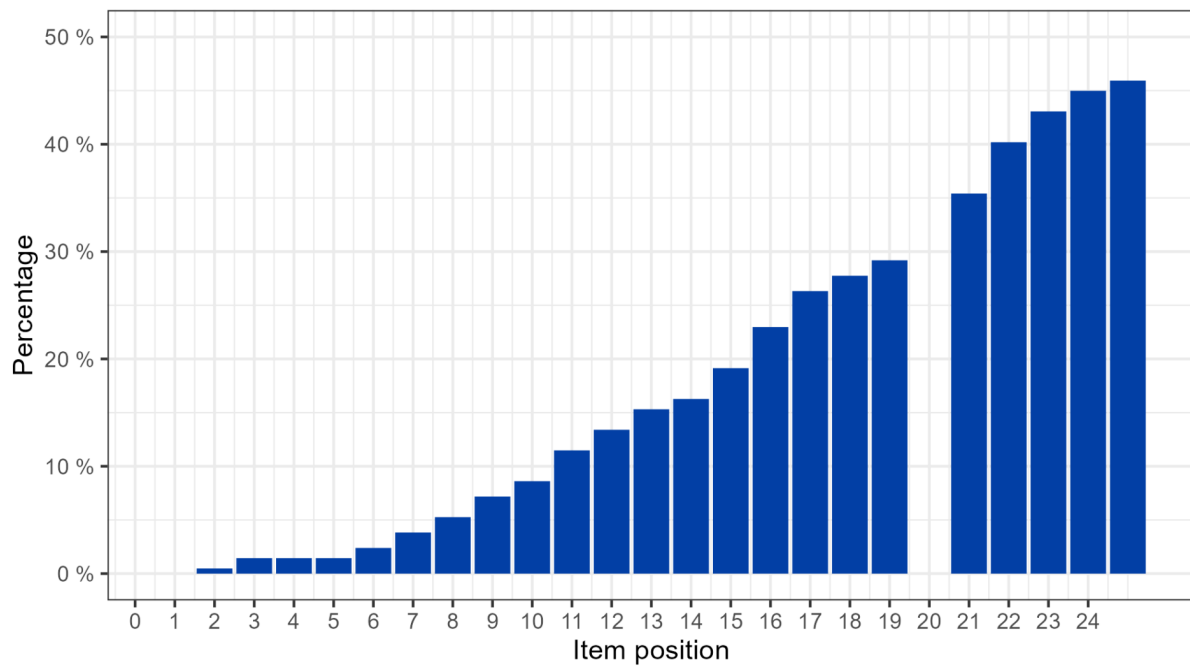
In contrast, there was a substantial number of items that respondents did not reach. The percentage of missing responses because participants did not reach an item was $Mdn = 14\%$. The percentage of respondents not reaching an item increased with the position of the item in the test (see Figure 5) up to 46%. Because a relatively large number of items was not reached by respondents, the total number of missing responses per item was also quite large. These varied between 5% and 49% ($Mdn = 20\%$).

5.1.2 Parameter Estimates

5.1.2.1 Item Parameters

The percentage of correct responses out of all valid responses for each item administered varied between 9.09% and 68.90% and had therefore an appropriate range for the sample (see Table 2).

The estimated item difficulties are given in Table 2. Item difficulties were estimated by constraining the mean of the ability distribution to zero. The estimated item difficulties ranged from -0.96 (vig50104_sc8g5_c) to 2.68 (vig50121_sc8g5_c) with a median of 1.33, thus covering a fairly wide range including both easy and difficult items. Because of the small sample size, the standard errors (SE) of the estimated item parameters were rather large with a median of 0.17 ($Min = 0.15$, $Max = 0.25$). The results reported in Table 2 are therefore rather imprecise.

Figure 5*Item Position not Reached for Verbal Reasoning*

Note. The item on position 20 was excluded due to an unsatisfactory item fit.

Table 2*Item Parameters for Verbal Reasoning*

Nr.	Item	<i>N</i>	Percentage correct	Difficulty	<i>SE</i>	WMNSQ	<i>t</i>	<i>r_{it}</i>	<i>Q₃</i>	Discr.
1	vig50101_sc8g5_c	209	61.72	-0.57	0.16	1.03	0.48	0.33	0.05	0.96
2	vig50102_sc8g5_c	209	56.94	-0.33	0.15	1.03	0.59	0.33	0.07	1.05
3	vig50103_sc8g5_c	209	12.92	2.24	0.22	1.05	0.39	0.19	0.06	0.65
4	vig50104_sc8g5_c	209	68.90	-0.96	0.16	0.94	-0.77	0.38	0.06	1.36
5	vig50105_sc8g5_c	209	53.11	-0.15	0.15	0.94	-0.98	0.41	0.09	1.56
6	vig50106_sc8g5_c	209	55.02	-0.24	0.15	0.90	-1.85	0.48	0.08	1.77
7	vig50107_sc8g5_c	209	27.27	1.18	0.17	1.09	1.11	0.22	0.07	0.66
8	vig50108_sc8g5_c	209	32.54	0.88	0.16	0.90	-1.52	0.48	0.08	1.77
9	vig50109_sc8g5_c	209	25.84	1.27	0.17	1.00	0.05	0.33	0.06	0.97
10	vig50110_sc8g5_c	209	46.41	0.18	0.15	0.92	-1.38	0.46	0.07	1.57
11	vig50111_sc8g5_c	209	27.27	1.18	0.17	0.96	-0.43	0.39	0.07	1.26

Nr.	Item	<i>N</i>	Percentage correct	Difficulty	<i>SE</i>	WMNSQ	<i>t</i>	<i>r_{it}</i>	<i>Q₃</i>	Discr.
12	vig50112_sc8g5_c	209	20.10	1.65	0.19	0.88	-1.22	0.48	0.09	2.04
13	vig50113_sc8g5_c	209	14.35	2.11	0.21	1.10	0.80	0.14	0.09	0.34
14	vig50114_sc8g5_c	209	33.01	0.86	0.16	0.91	-1.40	0.47	0.08	1.66
15	vig50115_sc8g5_c	209	16.27	1.94	0.20	1.03	0.30	0.26	0.08	0.68
16	vig50116_sc8g5_c	209	23.92	1.39	0.18	1.04	0.42	0.28	0.07	0.67
17	vig50117_sc8g5_c	209	26.79	1.21	0.17	1.00	0.05	0.34	0.07	0.87
18	vig50118_sc8g5_c	209	18.66	1.75	0.19	1.08	0.76	0.22	0.06	0.65
19	vig50119_sc8g5_c	209	16.75	1.90	0.20	1.14	1.17	0.13	0.07	0.33
20	vig50121_sc8g5_c	209	9.09	2.68	0.25	0.96	-0.14	0.25	0.06	0.94
21	vig50122_sc8g5_c	209	11.00	2.45	0.23	1.00	0.02	0.26	0.08	0.79
22	vig50123_sc8g5_c	209	15.31	2.02	0.20	1.11	0.91	0.12	0.07	0.30
23	vig50124_sc8g5_c	209	16.27	1.94	0.20	1.08	0.64	0.18	0.08	0.37
24	vig50125_sc8g5_c	209	15.79	1.98	0.20	0.93	-0.51	0.37	0.07	1.18

Note. *N* = Number of observed responses. Difficulty = Item difficulty. *SE* = Standard error of item parameter. WMNSQ = Weighted mean square error. *t* = *t*-value for WMNSQ. Discr. = Discrimination parameter of a two-parametric item response model. *r_{it}* = Corrected item-total correlation. *Q₃* = Average absolute residual correlation for item (Yen, 1983).

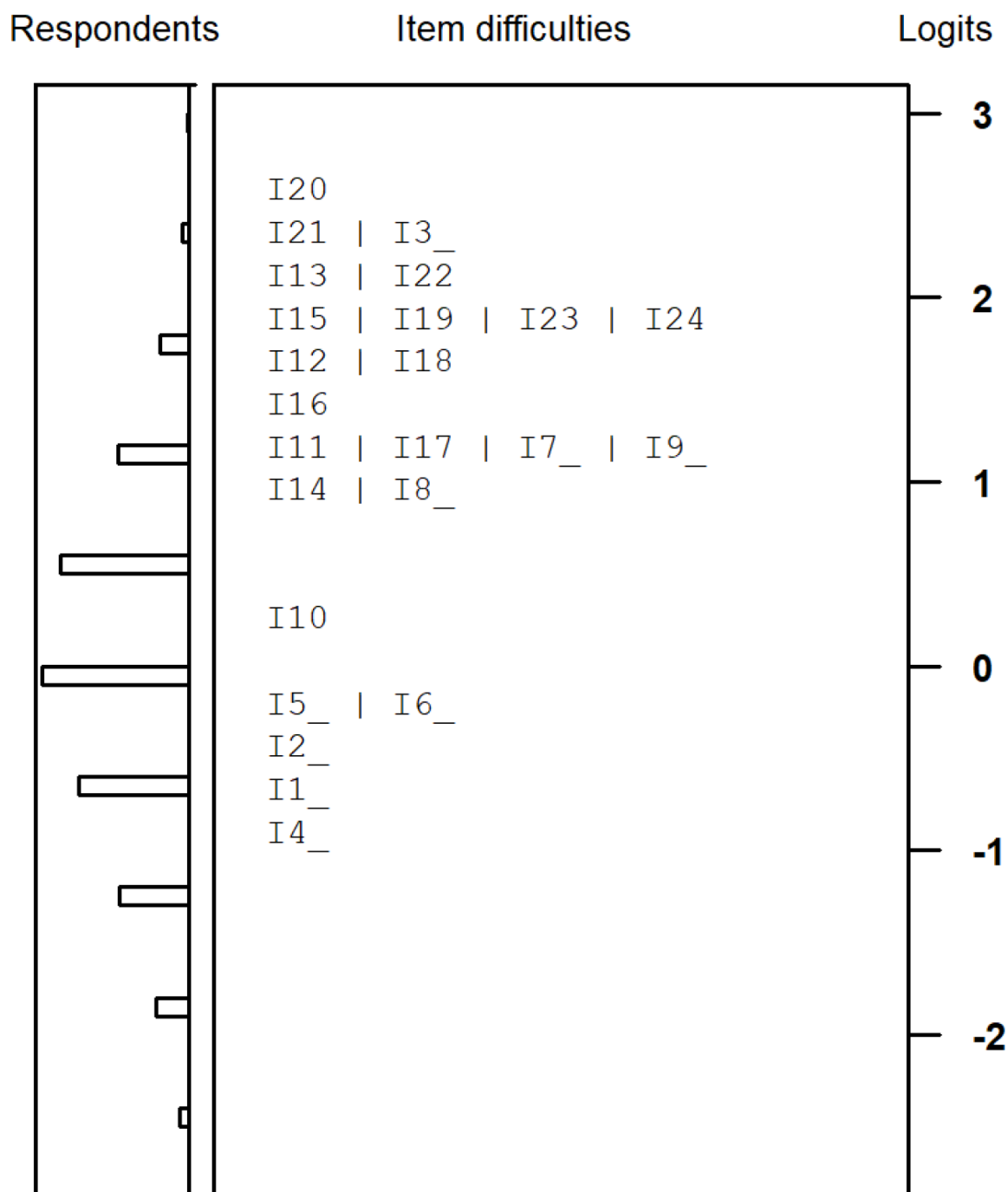
5.1.2.2 Test Targeting and Reliability

Test targeting focuses on comparing the item difficulties with the person abilities to evaluate the appropriateness of the test for the specific target population. In Figure 6, the item difficulties of the verbal reasoning items and the ability of the respondents are plotted on the same scale. The distribution of respondents' estimated abilities is plotted on the left, while the distribution of the item difficulties is plotted on the right.

The item difficulties ranged from -0.96 (vig50104_sc8g5_c) to 2.68 (vig50121_sc8g5_c) and thus covered a rather wide range. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 1.02, which implies a good differentiation between respondents. The reliability of the test (EAP/PV reliability = 0.77, WLE reliability = 0.74) was satisfactory. The median of the category threshold distribution was about 1.33 logits above the mean person ability distribution. Thus, on average the items were too difficult for the respondents. There were few easy items. As a result, medium to high proficiencies were measured relatively accurately, while lower ability estimates had larger standard errors of measurement.

Figure 6*Test Targeting for Verbal Reasoning*

Wright map



Note. The distribution of the person abilities in the sample is given on the left-hand side of the graph. The item difficulties are given on the right-hand side of the graph, with each number corresponding to the item numbers given in Table 2.

5.1.3 Quality of the test

5.1.3.1 Distractor Analyses

To investigate how well the distractors of the multiple-choice items performed, point-biserial correlations were calculated between each incorrect response (distractor) and the respondents' total correct scores for the remaining items. The median point-biserial correlation for the distractors fell at $-.10$ ($Min = -.38$, $Max = .28$) and thus indicated that on average the distractors worked well. However, particularly for very difficult items some distractors were disproportionately often selected by high-ability respondents. This led to point-biserial correlations for some distractors in several items that were positive (e.g., $r_{it} = .28$ for vig50123_sc8g5_c). These results suggest that some items might have been too difficult for the sample.

5.1.3.2 Item Fit

The evaluation of item fit was based on the final scaling model presented above. Overall, the item fit was satisfactory (see Table 2). The WMNSQ values ranged from 0.88 (vig50112_sc8g5_c) to 1.14 (vig50119_sc8g5_c), with a median of 1.00. There was no item with a WMSNQ greater than 1.15. Also, visual inspections of the empirical ICC for the item did not reveal any pronounced deviation from the expected ICCs. In addition, the correlations of the correct responses with the total test scores varied between 0.12 (vig50123_sc8g5_c) and 0.48 (vig50106_sc8g5_c, vig50108_sc8g5_c, vig50112_sc8g5_c), with a median of 0.33.

5.1.3.3 Differential Item Functioning

DIF was used to evaluate whether the measurement models were comparable between Starting Cohort 3 (Gnambs & Nusser, 2024) and Starting Cohort 8. The differences between the estimated item difficulties (on the logit scale) in the different groups are summarized in Table 3. The column "SC 3 vs. 8" reports the differences in item difficulties between Starting Cohort 3 and Starting Cohort 8; a positive value would indicate that the test was more difficult in Starting Cohort 3, whereas a negative value would indicate that the item was less difficult in Starting Cohort 3 than in Starting Cohort 8. In addition to examining DIF for each item, an overall test for DIF was performed by comparing models that allow for DIF with those that only estimate main effects.

There were 509 respondents from Starting Cohort 3 and 209 respondents from Starting Cohort 8. The small main effect in verbal reasoning between cohorts of 0.30 logits (Cohen's $d = 0.35$) showed slightly higher verbal reasoning abilities in Starting Cohort 3 than in Starting Cohort 8. There were 2 items with DIF effects greater than 0.60 (vig50108_sc8g5_c, vig50122_sc8g5_c). However, an overall test for DIF suggested a better fit for the model without DIF (AIC = 18,330, BIC = 18,449, number of parameters = 26) compared to a model that also acknowledged potential DIF (AIC = 18,330, BIC = 18,554, number of parameters = 49). These results indicate that for most items there was no pronounced DIF concerning the starting cohort that might have biased the parameter estimates.

Table 3*Differential Item Functioning for Verbal Reasoning*

Item	SC 3 vs. 8	Item	SC 3 vs. 8
vig50101_sc8g5_c	0.16 (0.19)	vig50113_sc8g5_c	0.05 (0.06)
vig50102_sc8g5_c	0.16 (0.19)	vig50114_sc8g5_c	0.22 (0.26)
vig50103_sc8g5_c	-0.26 (-0.31)	vig50115_sc8g5_c	-0.19 (-0.22)
vig50104_sc8g5_c	-0.01 (-0.01)	vig50116_sc8g5_c	-0.37 (-0.44)
vig50105_sc8g5_c	-0.01 (-0.01)	vig50117_sc8g5_c	-0.32 (-0.38)
vig50106_sc8g5_c	0.00 (0.00)	vig50118_sc8g5_c	0.45 (0.53)
vig50107_sc8g5_c	-0.05 (-0.06)	vig50119_sc8g5_c	0.06 (0.07)
vig50108_sc8g5_c	0.74 (0.87)	vig50121_sc8g5_c	0.19 (0.22)
vig50109_sc8g5_c	-0.06 (-0.07)	vig50122_sc8g5_c	-0.64 (-0.75)
vig50110_sc8g5_c	-0.06 (-0.07)	vig50123_sc8g5_c	-0.22 (-0.26)
vig50111_sc8g5_c	0.42 (0.50)	vig50124_sc8g5_c	-0.15 (-0.18)
vig50112_sc8g5_c	0.28 (0.33)	vig50125_sc8g5_c	0.39 (0.46)
Main effect (DIF model)	0.30 (0.35)	Main effect (Main effect model)	0.28 (0.34)

Note. SC = starting cohort. Raw differences between item difficulties (in logits) with standardized differences (Cohen's *d*) in parentheses.

5.1.3.4 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is that all item discrimination parameters are equal. To test this assumption, a two-parametric item response model (Birnbbaum, 1968) was fitted to the data to estimate the discrimination parameters. The estimated discrimination parameters differed moderately between items (see Table 2). The median discrimination parameter fell at 0.95 (*Min* = 0.30, *Max* = 2.04). Model fit indices suggested a comparable fit of the Rasch model (AIC = 5,025, BIC = 5,109, number of parameters = 25) and the two-parametric item response model (AIC = 4,998, BIC = 5,158, number of parameters = 48). For this reason and in line with the theoretical concepts underlying the test construction (see Heller & Perleth, 2000), the Rasch model was chosen as the scaling model in order to preserve the item weightings as intended in the theoretical framework.

5.1.3.5 Dimensionality

The dimensionality of the test was investigated by evaluating the correlations between the residuals of the Rasch model. The adjusted Q_3 statistics (see Table 2) were quite low (*Mdn* = 0.07) - the largest absolute residual correlation was 0.09 (vig50105_sc8g5_c,

vig50112_sc8g5_c, vig50113_sc8g5_c). Overall, these results indicate an essentially unidimensional test.

5.2 Nonverbal Reasoning

5.2.1 Missing Responses

5.2.1.1 Missing Responses per Person

The number of invalid responses per person is shown in Figure 7. The number of invalid responses was rather low with 96% of respondents having no invalid responses at all. Only about 1% of respondents had more than one invalid response.

Figure 7

Number of Invalid Responses for Nonverbal Reasoning

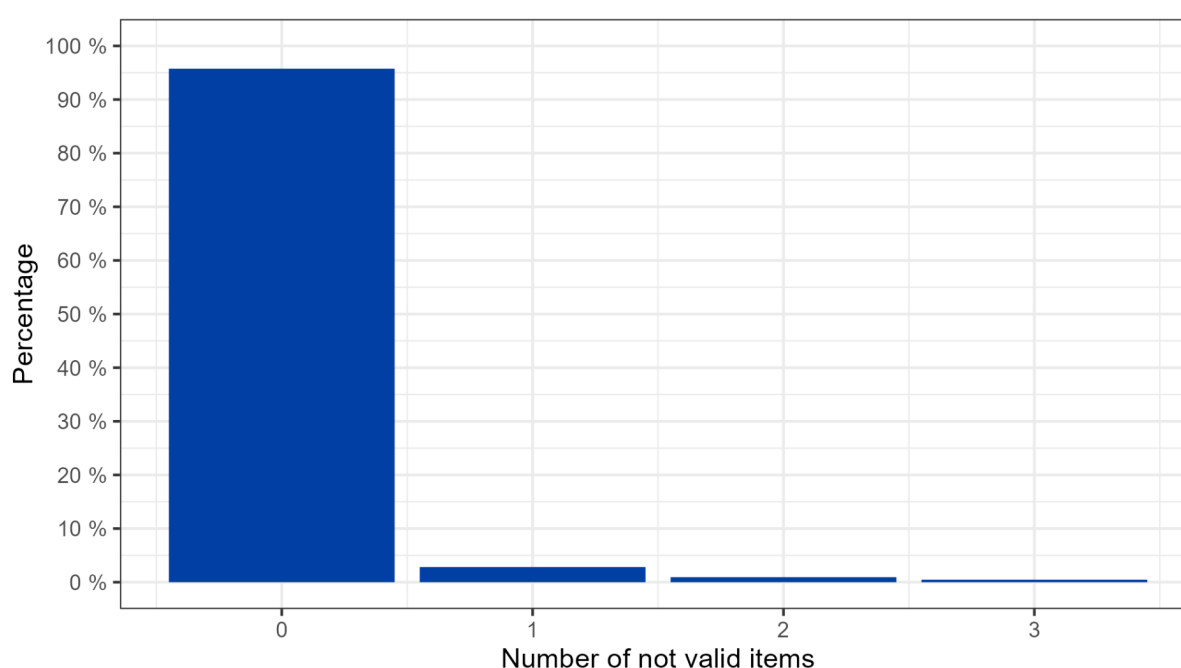


Figure 8 illustrates the number of missing responses when respondents omitted items. The figure shows approximately 78% of respondents had no omitted items, while 14% exhibited a single omitted item. However, less than 1% of respondents had five or more omitted items.

As shown in Figure 9, about 92% of respondents did not abort the test and were administered all items. Notably, the entire sample received at least 8 items and thus about a third of the total test.

The total number of missing responses, aggregated across invalid, omitted, and not reached missing responses for each respondent, is shown in Figure 10. Because most respondents reached the end of the test and did not omit any items, the total number of missing values was rather low. The median number of missing responses was 0 and about 70% of the respondents had no missing response at all. Only 4% of respondents had more than five missing responses, while 1% had missing responses for 12 or more items (i.e., almost half of the administered items).

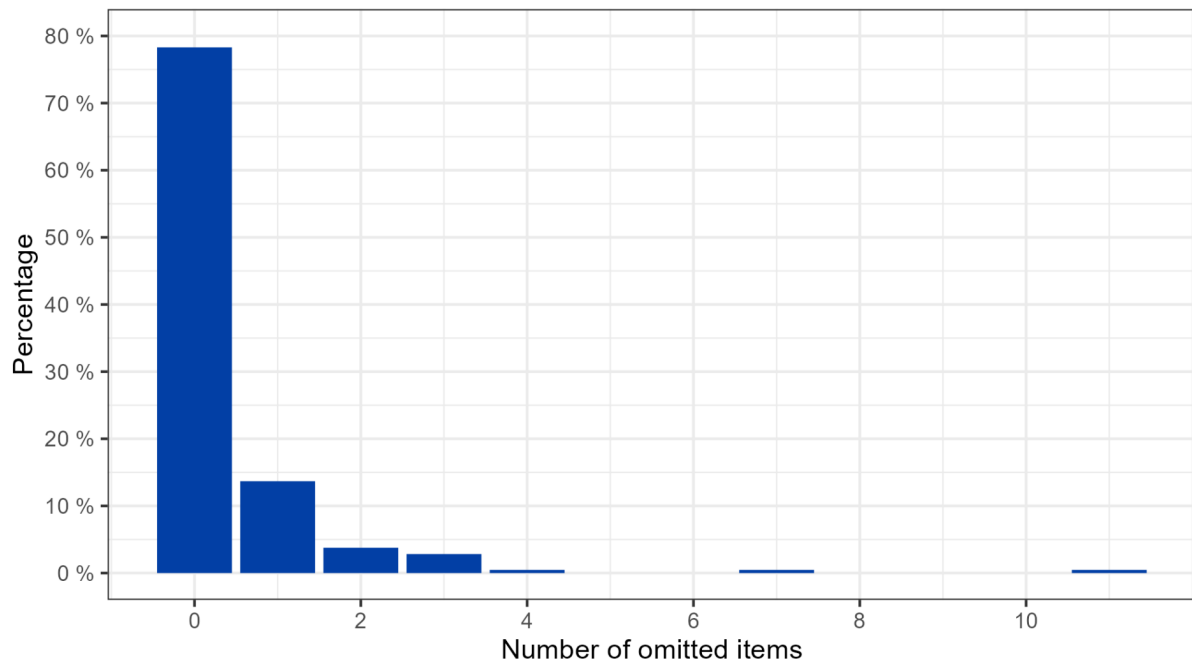
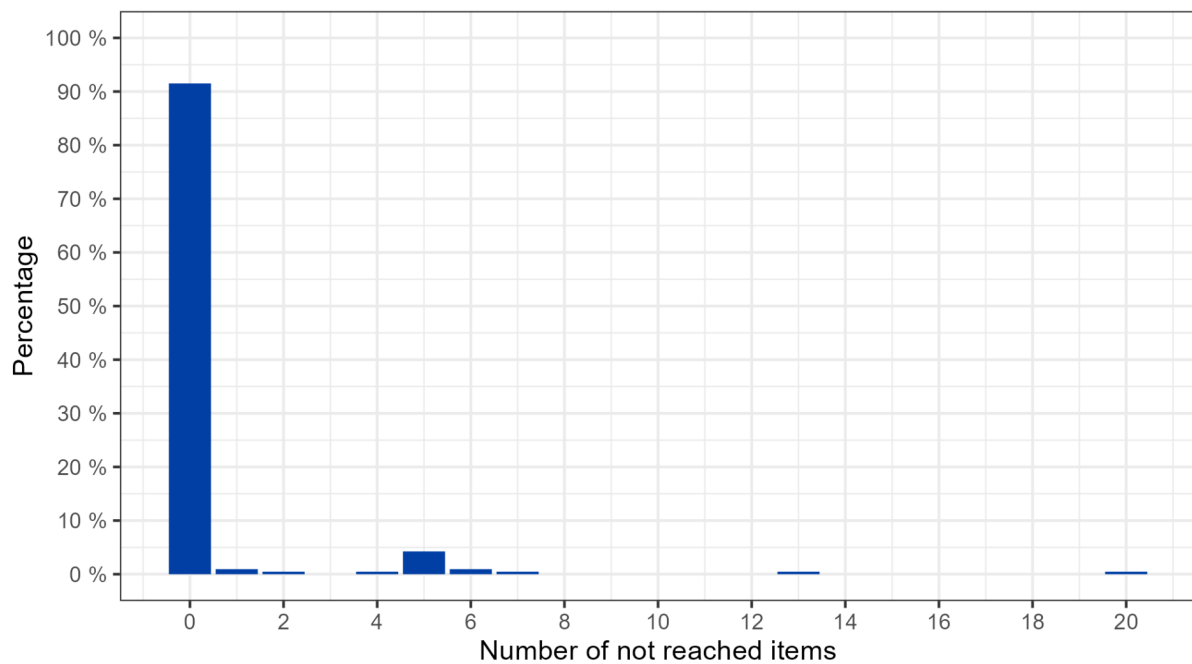
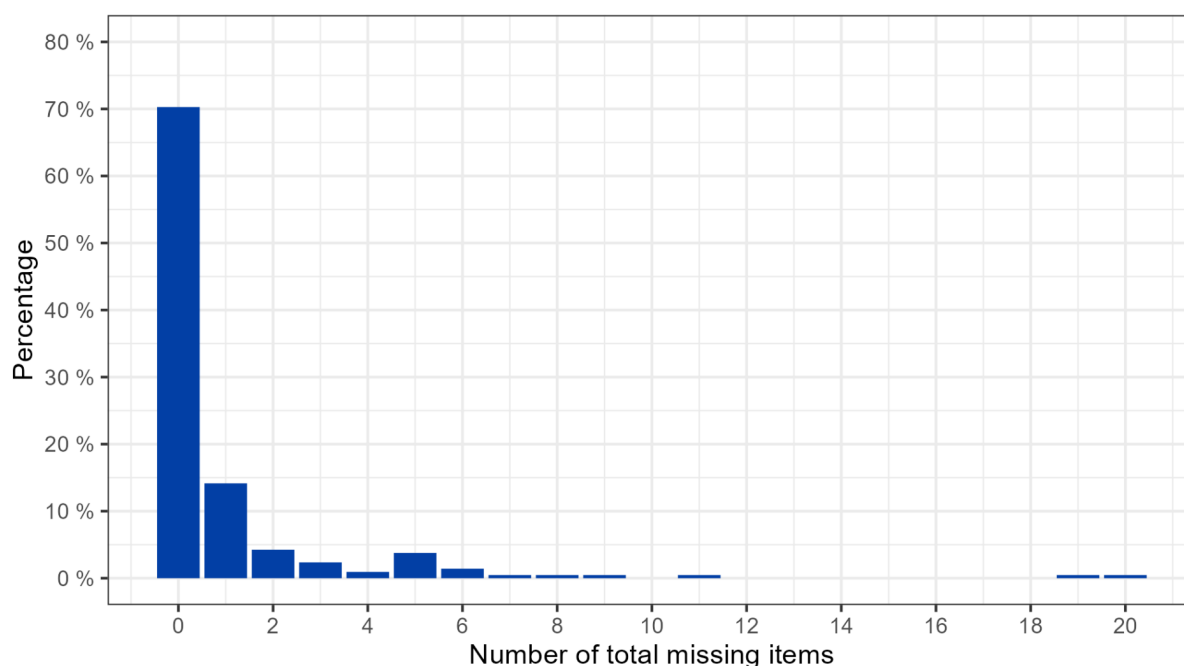
Figure 8*Number of Omitted Items for Nonverbal Reasoning***Figure 9***Number of Not-Reached Items for Nonverbal Reasoning*

Figure 10

Total Number of Missing Responses for Nonverbal Reasoning



Overall, there was a small amount of invalid missing responses or items that were omitted. Because most respondents reached the end of the test, also the total number of missing responses was rather small.

5.2.1.2 Missing Responses per Item

Table 4 provides information on the occurrence of different types of missing responses per item. Overall, the number of respondents that omitted an item was acceptable. The percentage of omitted responses per item varied between 0% and 7% ($Mdn = 1\%$). In addition, most items had relatively few invalid responses. The percentage of invalid responses varied across items between 0% and 1% ($Mdn = 0\%$).

The percentage of missing responses because participants did not reach an item was $Mdn = 1\%$. The percentage of respondents not reaching an item slightly increased with the position of the item in the test (see Figure 11) up to 8%. Because of the few invalid, omitted, and not-reached responses, the total number of missing responses per item was also small. These varied between 0% and 10% ($Mdn = 2\%$).

Table 4

Percentage of Missing Responses by Item for Nonverbal Reasoning

Nr.	Item	Pos.	N	OM	NR	NV
1	nig50101_sc8g5_c	1	208	1.42	0.00	0.47
2	nig50102_sc8g5_c	2	208	0.94	0.00	0.94

Nr.	Item	Pos.	N	OM	NR	NV
3	nig50103_sc8g5_c	3	207	2.36	0.00	0.00
4	nig50104_sc8g5_c	4	208	1.42	0.00	0.47
5	nig50105_sc8g5_c	5	209	1.42	0.00	0.00
6	nig50106_sc8g5_c	6	208	1.42	0.47	0.00
7	nig50107_sc8g5_c	7	195	7.08	0.47	0.47
8	nig50108_sc8g5_c	8	207	1.42	0.47	0.47
9	nig50109_sc8g5_c	9	211	0.00	0.47	0.00
10	nig50110_sc8g5_c	10	210	0.00	0.47	0.47
11	nig50111_sc8g5_c	11	207	1.89	0.47	0.00
12	nig50112_sc8g5_c	12	207	1.42	0.47	0.47
13	nig50113_sc8g5_c	13	208	0.94	0.94	0.00
14	nig50114_sc8g5_c	14	205	1.42	0.94	0.94
15	nig50115_sc8g5_c	15	209	0.47	0.94	0.00
16	nig50116_sc8g5_c	16	201	4.25	0.94	0.00
17	nig50117_sc8g5_c	17	207	0.94	0.94	0.47
18	nig50118_sc8g5_c	18	204	2.83	0.94	0.00
19	nig50119_sc8g5_c	19	207	0.94	1.42	0.00
20	nig50120_sc8g5_c	20	206	0.47	2.36	0.00
21	nig50121_sc8g5_c	21	194	1.89	6.60	0.00
22	nig50122_sc8g5_c	22	195	0.94	7.08	0.00
23	nig50123_sc8g5_c	23	190	3.30	7.08	0.00
24	nig50124_sc8g5_c	24	192	0.94	7.55	0.94
25	nig50125_sc8g5_c	25	194	0.00	8.49	0.00

Note. Pos. = Item position within the test. N = Number of valid responses. NR = Percentage of respondents that did not reach an item. OM = Percentage of omitted responses. NV = Percentage of invalid responses.

5.2.2 Parameter Estimates

5.2.2.1 Item Parameters

The percentage of correct responses out of all valid responses for each item administered varied between 8.96% and 95.28% and had therefore an appropriate range for the sample (see

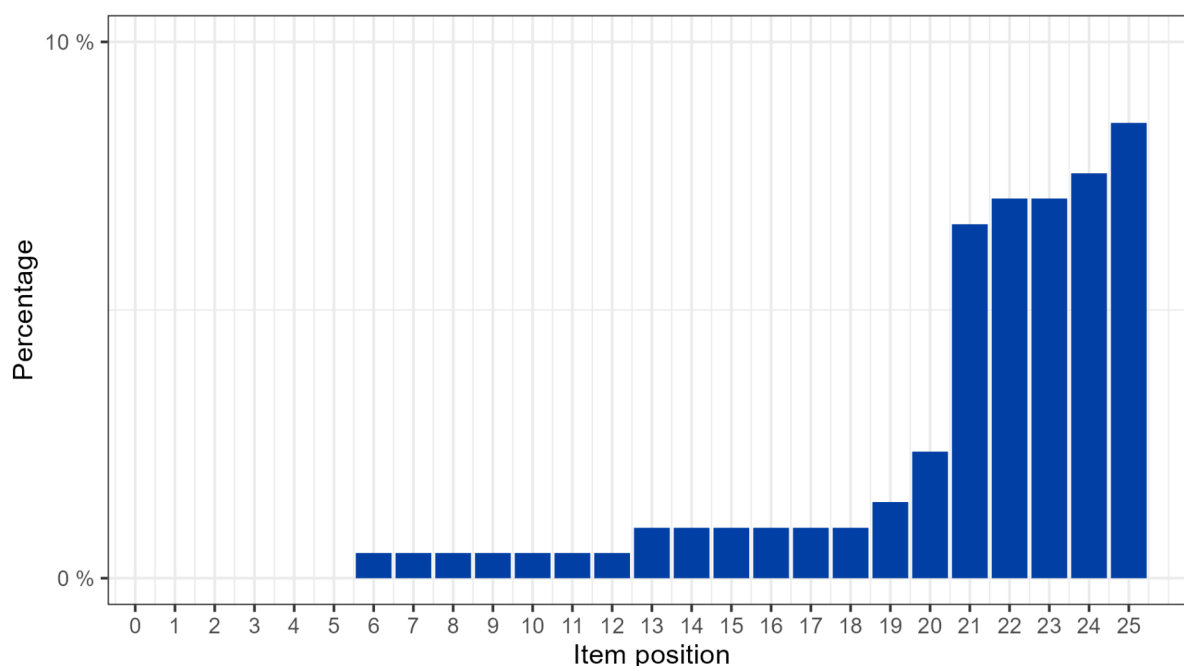
Figure 11*Item Position not Reached for Nonverbal Reasoning*

Table 5). Because there was a non-negligible amount of missing responses, these probabilities cannot be interpreted as an index of item difficulty.

The estimated item difficulties are given in Table 5. Item difficulties were estimated by constraining the mean of the ability distribution to zero. The estimated item difficulties ranged from -3.47 (nig50101_sc8g5_c) to 2.75 (nig50125_sc8g5_c) with a median of -1.43, thus covering a fairly wide range including both easy and difficult items. Because of the small sample size, the standard errors (*SE*) of the estimated item parameters were rather large with a median of 0.18 (*Min* = 0.15, *Max* = 0.33). The results reported in Table 5 are therefore rather imprecise.

Table 5*Item Parameters for Nonverbal Reasoning*

Nr.	Item	<i>N</i>	Percentage correct	Difficulty	<i>SE</i>	WMNSQ	<i>t</i>	<i>r_{it}</i>	<i>Q₃</i>	Discr.
1	nig50101_sc8g5_c	212	95.28	-3.47	0.33	1.01	0.12	0.22	0.08	1.39
2	nig50102_sc8g5_c	212	87.74	-2.34	0.22	1.05	0.40	0.23	0.07	0.93
3	nig50103_sc8g5_c	212	84.43	-2.03	0.20	0.98	-0.08	0.32	0.06	1.28
4	nig50104_sc8g5_c	212	87.74	-2.34	0.22	1.07	0.53	0.18	0.07	0.54
5	nig50105_sc8g5_c	212	82.55	-1.87	0.19	1.08	0.76	0.22	0.07	0.46
6	nig50106_sc8g5_c	212	94.34	-3.26	0.31	0.96	-0.07	0.22	0.07	1.00
7	nig50107_sc8g5_c	212	81.60	-1.80	0.19	1.06	0.54	0.24	0.09	0.57

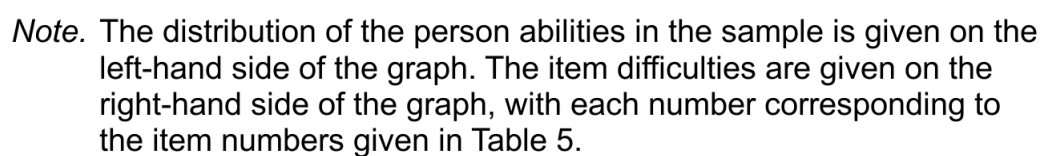
Nr.	Item	<i>N</i>	Percentage correct	Difficulty	<i>SE</i>	WMNSQ	<i>t</i>	r_{it}	Q_3	Discr.
8	nig50108_sc8g5_c	212	88.68	-2.44	0.23	1.02	0.17	0.23	0.07	0.77
9	nig50109_sc8g5_c	212	81.60	-1.80	0.19	1.14	1.24	0.13	0.07	0.35
10	nig50110_sc8g5_c	212	82.08	-1.84	0.19	0.96	-0.29	0.37	0.09	1.67
11	nig50111_sc8g5_c	212	68.40	-0.95	0.16	0.85	-2.28	0.55	0.12	2.82
12	nig50112_sc8g5_c	212	64.62	-0.75	0.16	1.02	0.29	0.35	0.06	0.81
13	nig50113_sc8g5_c	212	76.42	-1.43	0.18	0.98	-0.14	0.36	0.06	1.10
14	nig50114_sc8g5_c	212	69.81	-1.03	0.16	0.81	-2.69	0.59	0.14	3.85
15	nig50115_sc8g5_c	212	78.30	-1.56	0.18	1.09	0.96	0.21	0.07	0.52
16	nig50116_sc8g5_c	212	48.58	0.06	0.15	1.06	1.07	0.32	0.06	0.66
17	nig50117_sc8g5_c	212	45.28	0.22	0.15	1.01	0.24	0.37	0.08	1.19
18	nig50118_sc8g5_c	212	78.30	-1.56	0.18	1.20	1.98	0.07	0.07	0.11
19	nig50119_sc8g5_c	212	49.06	0.04	0.15	0.90	-1.76	0.50	0.08	1.69
20	nig50120_sc8g5_c	212	61.79	-0.60	0.16	0.96	-0.67	0.43	0.06	1.37
21	nig50121_sc8g5_c	212	54.72	-0.24	0.15	0.88	-2.07	0.51	0.07	1.59
22	nig50122_sc8g5_c	212	57.08	-0.36	0.15	0.85	-2.71	0.56	0.09	2.14
23	nig50123_sc8g5_c	212	43.87	0.30	0.15	0.94	-0.95	0.44	0.05	1.34
24	nig50124_sc8g5_c	212	47.17	0.13	0.15	1.11	1.89	0.25	0.05	0.68
25	nig50125_sc8g5_c	212	8.96	2.75	0.25	1.07	0.41	0.14	0.04	0.60

Note. *N* = Number of observed responses. Difficulty = Item difficulty. *SE* = Standard error of item parameter. WMNSQ = Weighted mean square error. *t* = *t*-value for WMNSQ. Discr. = Discrimination parameter of a two-parametric item response model. r_{it} = Corrected item-total correlation. Q_3 = Average absolute residual correlation for item (Yen, 1983).

5.2.2.2 Test Targeting and Reliability

Figure 12 plots the item difficulties of the nonverbal reasoning items and the ability of the respondents. The distribution of respondents' estimated abilities is plotted on the left, while the distribution of the item difficulties is plotted on the right.

The item difficulties ranged from -3.47 (nig50101_sc8g5_c) to 2.75 (nig50125_sc8g5_c) and thus covered a rather wide range. The mean of the ability distribution was constrained to be zero. The variance was estimated to be 1.12, which implies a good differentiation between respondents. The reliability of the test (EAP/PV reliability = 0.79, WLE reliability = 0.76) was satisfactory. The median of the category threshold distribution was about -1.44 logits below the mean person ability distribution. Thus, on average the items were too easy for the respondents. There were few difficult items. As a result, low to medium proficiencies were measured rela-



tively accurately, while higher ability estimates had larger standard errors of measurement.

5.2.3 Quality of the test

5.2.3.1 Distractor Analyses

To investigate how well the distractors of the multiple-choice items performed, point-biserial correlations were calculated between each incorrect response (distractor) and the respondents' total correct scores for the remaining items. The median point-biserial correlation for the distractors fell at $-.09$ ($Min = -.35$, $Max = .15$) and thus indicated that on average the distractors worked well. Again, some items included distractors that were disproportionately often selected by high-ability respondents. This led to point-biserial correlations for some distractors that were positive (e.g., $r_{it} = .15$ for nig50117_sc8g5_c). However, this was observed for rather few items.

5.2.3.2 Item Fit

The evaluation of item fit was based on the final scaling model presented above. Overall, the item fit was satisfactory (see Table 5). The WMNSQ values ranged from 0.81 (nig50114_sc8g5_c) to 1.20 (nig50118_sc8g5_c), with a median of 1.01. There was 1 item with a WMNSQ greater than 1.15 (nig50118_sc8g5_c), while none of them had a WMNSQ greater than 1.20. Also, visual inspections of the empirical ICC for the item did not reveal any pronounced deviation from the expected ICCs. In addition, the correlations of the correct responses with the total test scores varied between 0.07 (nig50118_sc8g5_c) and 0.59 (nig50114_sc8g5_c), with a median of 0.32.

5.2.3.3 Differential Item Functioning

DIF was used to evaluate whether the measurement models were comparable between Starting Cohort 3 (Gnambs & Nusser, 2024) and Starting Cohort 8. The differences between the estimated item difficulties (on the logit scale) in the different groups are summarized in Table 6.

There were 510 respondents from Starting Cohort 3 and 212 respondents from Starting Cohort 8. The small main effect in nonverbal reasoning between cohorts of -0.49 logits (Cohen's $d = -0.45$) showed higher nonverbal reasoning abilities in Starting Cohort 8 than in Starting Cohort 3. Non-negligible DIF effects exceeding 0.60 logits were observed for 6 items, with 3 of them exhibiting DIF greater than 1.00 logits (nig50111_sc8g5_c, nig50115_sc8g5_c, nig50125_sc8g5_c). However, an overall test for DIF suggested a comparable fit for the model without DIF (AIC = 18,881, BIC = 19,004, number of parameters = 27) as a model that also acknowledged potential DIF (AIC = 18,774, BIC = 19,007, number of parameters = 51). These results suggest that the DIF effects were estimated rather imprecisely and, overall, DIF did not seem to be pronounced. Therefore, it is unlikely that systematic DIF concerning the starting cohort might bias cross-cohort comparisons.

Table 6*Differential Item Functioning for Nonverbal Reasoning*

Item	SC 3 vs. 8	Item	SC 3 vs. 8
nig50101_sc8g5_c	0.15 (0.14)	nig50114_sc8g5_c	-0.68 (-0.62)
nig50102_sc8g5_c	-0.21 (-0.19)	nig50115_sc8g5_c	1.12 (1.03)
nig50103_sc8g5_c	-0.32 (-0.29)	nig50116_sc8g5_c	-0.04 (-0.04)
nig50104_sc8g5_c	0.23 (0.21)	nig50117_sc8g5_c	0.17 (0.16)
nig50105_sc8g5_c	0.68 (0.62)	nig50118_sc8g5_c	0.33 (0.30)
nig50106_sc8g5_c	0.25 (0.23)	nig50119_sc8g5_c	-0.28 (-0.26)
nig50107_sc8g5_c	0.33 (0.30)	nig50120_sc8g5_c	0.13 (0.12)
nig50108_sc8g5_c	0.63 (0.58)	nig50121_sc8g5_c	0.07 (0.06)
nig50109_sc8g5_c	0.14 (0.13)	nig50122_sc8g5_c	-0.43 (-0.39)
nig50110_sc8g5_c	0.20 (0.18)	nig50123_sc8g5_c	-0.33 (-0.30)
nig50111_sc8g5_c	-1.33 (-1.22)	nig50124_sc8g5_c	0.11 (0.10)
nig50112_sc8g5_c	0.36 (0.33)	nig50125_sc8g5_c	1.38 (1.27)
nig50113_sc8g5_c	0.11 (0.10)		
Main effect (DIF model)	-0.49 (-0.45)	Main effect (Main effect model)	-0.48 (-0.45)

Note. SC = starting cohort. Raw differences between item difficulties (in logits) with standardized differences (Cohen's *d*) in parentheses.

5.2.3.4 Rasch-homogeneity

An essential assumption of the Rasch (1960) model is that all item discrimination parameters are equal. To test this assumption, a two-parametric item response model (Birnbaum, 1968) was fitted to the data to estimate the discrimination parameters. The estimated discrimination parameters differed moderately between items (see Table 5). The median discrimination parameter fell at 1.00 (*Min* = 0.11, *Max* = 3.85). Model fit indices suggested a comparable fit of the Rasch model (AIC = 5,164, BIC = 5,252, number of parameters = 26) and the two-parametric item response model (AIC = 5,080, BIC = 5,248, number of parameters = 50). For this reason and in line with the theoretical concepts underlying the test construction (see Heller & Perleth, 2000), the Rasch model was chosen as the scaling model in order to preserve the item weightings as intended in the theoretical framework.

5.2.3.5 Dimensionality

The dimensionality of the test was investigated by evaluating the correlations between the residuals of the Rasch model. The adjusted Q_3 statistics (see Table 5) were quite low (*Mdn* =

0.07) - the largest absolute residual correlation was 0.14 (nig50114_sc8g5_c). Overall, these results indicate an essentially unidimensional test.

6. Discussion

The analyses in the previous sections provide information on the quality of two subscales of the *Cognitive Abilities Test* (Heller & Perleth, 2000) measuring verbal and nonverbal reasoning abilities that were administered in special schools of Starting Cohort 8 of the NEPS. Different types of missing responses were examined, item fit statistics and item characteristic curves were evaluated, and item discriminations were investigated. Further quality inspections were conducted by testing Rasch-homogeneity. Several criteria indicated a good fit of the items. Also, the number of missing values was not excessive and fell in line with missing rates in comparable samples.

The verbal reasoning subtest was better targeted at medium- to high-ability respondents because few items were available that could adequately discriminate in the lower proficiency range. In contrast, the nonverbal reasoning subtest was better targeted at low- to medium-ability respondents because it lacked difficult items. As a result, ability estimates for the verbal subtest were more accurate for high-performing respondents, while for the nonverbal subtest, the respective ability estimates were more accurate for low-ability respondents. Overall, however, the subtests had satisfactory reliabilities and adequately discriminated between respondents.

Analyses of differential item functioning identified several items that had different difficulties in Starting Cohort 3 and Starting Cohort 8, more so, in the nonverbal reasoning subtest than the verbal reasoning subtest. However, overall model evaluations in both cases favored models that ignored potential DIF effects. Therefore, it can be reasoned that the point estimates of DIF were rather imprecise resulting from the small sample sizes.

In conclusion, the two scales had satisfactory psychometric properties that allowed the estimation of unidimensional reasoning scores for the participants in special schools.

7. Data in the Scientific Use Files

7.1 Naming Conventions

The SUF for the starting cohort contains 25 items for each subtest that are scored dichotomously with 0 indicating an incorrect response and 1 indicating a correct response. Further details on the naming conventions of the variables can be found in Fuß et al. (2025).

7.2 Cognitive Ability Scores

In the SUF, manifest cognitive ability scores for both subtests are provided in the form of WLEs (vig5_sc1, nig5_sc1) including their respective standard errors (vig5_sc2, nig5_sc2). For the verbal subtest scores, only 24 items were considered in these analyses (see Section 2). The person abilities were estimated in a single model including respondents from special schools in Starting Cohort 3 (Gnambs & Nusser, 2024) and Starting Cohort 8. As a result, the WLEs can be used to compare the two cohorts. In addition, the SUF also provides sum scores (vig5_sc3, nig5_sc3) based on 25 items for each subtest (as suggested by the test authors, see Heller &

Perleth, 2000). No person scores were estimated for respondents who did not participate in the test or who did not provide enough valid responses. The values of the scores for these persons are denoted as not-determinable missing values.

References

- Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6), 716–722. <https://doi.org/10.1109/TAC.1974.1100705>
- Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick (Eds.), *Statistical theories of mental test scores* (pp. 397–479). Addison-Wesley.
- Fuß, D., Gnambs, T., Lockl, K., Attig, M., & Nusser, L. (2025). *Competence data in NEPS: Overview of measures and variable naming conventions (Starting Cohorts 1 to 6)*. Leibniz Institute for Educational Trajectories.
- Gnambs, T. (2025). *NEPS Technical Report for Reading: Scaling Results of Starting Cohort 8 in Grade 5 in Special Schools* (NEPS Survey Paper 118). Leibniz Institute for Educational Trajectories. <https://doi.org/10.5157/NEPS:SP118:1.0>
- Gnambs, T., & Nusser, L. (2024). Out-of-level cognitive testing of children with special educational needs. *European Journal of Psychological Assessment*, 40(1), 40–45. <https://doi.org/10.1027/1015-5759/a000736>
- Heller, K. A., & Perleth, C. (2000). *Kognitiver Fähigkeitstest für 4. bis 12. Klassen, Revision*. Beltz.
- Pohl, S., & Carstensen, C. H. (2012). *NEPS Technical Report – Scaling the data of the competence tests* (NEPS Working Paper 14). University of Bamberg, National Educational Panel Study. <https://doi.org/10.5157/NEPS:WP14:1.0>
- R Core Team. (2024). *R: A language and environment for statistical computing* (Version 4.4.1) [Computer software]. R Foundation for Statistical Computing. <https://www.R-project.org/>
- Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests*. Nielsen & Lydiche.
- Robitzsch, A., Kiefert, T., & Wu, M. (2024). *TAM: Test analysis modules* (Version 4.2-21) [Computer software]. <https://doi.org/10.32614/CRAN.package.TAM>
- Schwarz, G. E. (1978). Estimating the dimension of a model. *Annals of Statistics*, 6(2), 461–464. <https://doi.org/10.1214/aos/1176344136>
- Thorndike, R. L., & Hagen, E. (1971). *Cognitive abilities test*. Houghton Mifflin Company.
- Weinert, S., Artelt, C., Prenzel, M., Senkbeil, M., Ehmke, T., & Carstensen, C. H. (2011). Development of competencies across the life span. In H.-P. Blossfeld & H.-G. Roßbach (Eds.), *Zeitschrift für Erziehungswissenschaften*, 14. *Education as a lifelong process: The German National Educational Panel Study (NEPS)* (pp. 67–86). VS Verlag für Sozialwissenschaften. <https://doi.org/10.1007/s11618-011-0182-7>
- Yen, W. M. (1984). Effects of local item dependence on the fit and equating performance of the three-parameter logistic model. *Applied Psychological Measurement*, 8(2), 125–145. <https://doi.org/10.1177/014662168400800201>
- Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement*, 30(3), 187–213. <https://doi.org/10.1111/j.1745-3984.1993.tb00423.x>